DOCUMENT RESUME

ED 244 738 PS 014 298

AUTHOR Glass, Gene V.; Camilli, Gregory A.
TITLE "Follow Through" Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 1 Feb 81
NOTE 34p.
PUB TYPE Viewpoints (120) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Early Childhood Education; *Evaluation Methods; *Evaluation Needs; Federal Government; Government Role; *Program Evaluation; *Research Problems
IDENTIFIERS Chronbachs 95 Theses; *Project Follow Through; Qualitative Research; *Quantitative Research

ABSTRACT

Two questions are addressed in this document: What is worth knowing about Project Follow Through? and, How should the National Institute of Education (NIE) evaluate the Follow Through program? Discussion of the first question focuses on findings of past Follow Through evaluations, problems associated with the use of experimental design and statistics, and prospects for discovering new knowledge about the program. With respect to the second question, it is suggested that NIE should conduct evaluation emphasizing an ethnographic, principally descriptive case-study approach to enable informed choice by those involved in the program. The discussion is based on the following assumptions: (1) Past evaluations of Follow Through have been quantitative, experimental approaches to deriving value judgments; (2) The deficiencies of quantitative, experimental evaluation approaches are so thorough and irreparable as to disqualify their use; (3) There are probably at most a half-dozen important approaches to teaching children, and these are already well-represented in existing Follow Through models; and (4) The audience for Follow Through evaluations is an audience of teachers to whom appeals to the need for accountability for public funds or the rationality of science are largely irrelevant. Appended to the discussion are Chronbach's 95 theses about the proper roles, methods, and uses of evaluation. Theses running counter to a federal model of program evaluation are asterisked. (RH)

1 February 1981

"Follow Through" Evaluation

Gene V Glass

Gregory A. Camilli

University of Colorado-Boulder

What is worth knowing about Follow Through in light of 1) past studies, 2) current technical capability of measurement and experimental analysis, and 3) the impact of such knowledge once found? How should the National Institute of Education fulfill its new responsibility to evaluate Follow Through? These are questions that may lead to unanticipated and uncomfortable answers (unanticipated by NIE, perhaps, and uncomfortable to the technical experts -- psychologists, statisticians, and the like -- with whom NIE sees itself allied in the manufacture of knowledge).

Our theses before our arguments:

1. Past evaluations of Follow Through were quantitative and experimental. They created more dissent than consensus; they changed few minds. We will not endorse, for example, what those who wrote the October 1, 1980, NIE planning document believe was proved by the SRI/Abt evaluation, namely, that "Models that emphasized basic skills produced more gains in those areas and in self concept than other models" (p. 3). "Models" of compensatory education are minor influences in pupils' development. Far more important in children's growth are their native endowment, their health, how their parents and siblings treat them, and other influences not controlled by schools.

2. The deficiencies of quantitative, experimental evaluation approaches like those continually pressed on the federal government (e.g., see the "Planning Information Study for Future Follow Through Experiments" produced by a team at the University of Georgia, 25 May 1979) are thorough and irreparable. The problem lies less with experimental designs for assessing causal impact (the state of this art is adequate) than with the impossibility of translating complex, subtle and vague notions of child development and education into tests for mass administration. And the notion that this problem will be resolved with matrix sampling, logistic item models, factor analysis and the like, merely betrays a shallow understanding of the real difficulties faced by those who wish to "quantify" education.

3. There are not 22 models of compensatory education (though there are clearly that many and more groups of professionals who can put together a "program" and write a grant proposal), nor are there dozens of "new" models waiting to be discovered. There are probably at most a half-dozen importantly different approaches to teaching children and these are already well-represented in existing Follow Through models. A gimmick isn't a model; nor is a model an enthusiasm of a researcher recently emerged from the laboratory with a scientific solution to the problem of why some children learn uncertainly or not at all.

4. The proper audience for Follow Through evaluations is teachers -- ignoring for the moment the bought-audience of scholars and researchers who write and read Follow Through reports only when they are paid to do so. Teachers do not heed the statistical findings of experiments

when deciding how best to educate children. (Nor do you, reader, study experiments that tell you how best to evaluate Follow Through.) They decide such matters on the basis of complicated public and private understandings, beliefs, motives and wishes. They have the right and good reasons so to decide, and neither that right nor those reasons are changed one whit by appeals to the need for accountability for public funds or the rationality of science.

WHAT IS WORTH KNOWING ABOUT FOLLOW THROUGH?

1. Findings of Past Follow Through Evaluations.

During the 1970's, the US Office of Education spent about $20 million evaluating Follow-Through. The USOE/SRI/ABT evaluation of Follow Through has been judged defective in many respects (House, Glass, McLean & Walker, 1978). Follow Through models were classified in misleading ways. Outcome measures were adequate only for assessing the simplest mechanical skills. Attempts to measure progress toward other than the most narrow academic goals were unsuccessful. The evaluation was unfair to models that concentrated on goals beyond simple academic performance. The Follow Through evaluation proved only that differences in models even on the few simple outcomes measured were trivially small in comparison to the large differences among sites for the same model. The tiny differences among models that did exist skipped around perplexingly depending on how one resolved any one of several nuances of statistical analysis (Camilli, 1980; Bereiter and Kurland, 1980).

House and his colleagues (1978) drew these lessons from the costly experience.

> "The truth about Follow Through is complex. No simple answer to the problem of educating disadvantaged students has been found. Even with the narrow outcome measures employed, models that worked well in one town worked poorly in another. Unique features of the local setting had more effect on test scores than did the models. This does not mean that federal programs are useless or inappropriate for pursuing national objectives; however, many of the most significant factors affecting educational achievement lie outside the control of federal officials. Educational policy makers should expect that the same program may have quite different effects in different settings." (p. 156) "Enough experience exists to suggest that these massive experiments with narrow outcome measures are bad investments. The results are highly equivocal, and groups such as sponsors and parents feel excluded, even abused, because their goals and interests are not represented in the evaluation. A pluralistic society requires a variety of evaluation criteria and approaches. Groups that are significantly affected by an evaluation must have their interests reflected in the valuative criteria or they will perceive the results as illegitimate." (p. 158)
>
> "An even broader question is whether the federal government should be advocating particular models of instruction at all. On the basis of previous experience and this evaluation, we think not. Government advocacy of particular instructional models assumes both the feasibility of wide implementation and the similarity of effects in different locales. However flawed, this evaluation does suggest these assumptions are contrary to fact. When combined with the experience of other government programs, the evidence is strong that educational improvement strategies that acknowledge local circumstances will be far more effective in the long run." (p. 158)

In much of the debate about the effectiveness of Follow Through, the role of the model sponsors has been neglected. Though some sponsors have strongly contested the results of the Abt evaluation (Stebbins et al., 1977), we believe the interests of the sponsors have not been sufficiently clear. Recently, however, they published a report in which their stake in the Follow Through endeavor is aired (Hodges et al., 1980). We will not review this document in depth here, but will discuss a few of their recommendations.

Regarding the outcome of the Follow Through experiment, the model sponsors observed that

> These (i.e., Follow Through) successes are impressive, but they are not sufficient. Promising approaches to teaching disadvantaged children have been demonstrated, but still too little is known about how to make schooling effective and pleasant for large numbers of children. Many children are still performing well below their potential. The long term effects of intervention in the primary grades are not known. The effectiveness of the components of several instructional approaches was not revealed. Many questions remain about how to insure that good instructional practices become widespread. The impediments to implementation of systematic educational approaches are still present and not yet fully understood. (p. 73)

The model sponsors stated further that

> The implementation problems must be solved since it is apparent that many people are unhappy with the way the schools presently serve economically disadvantaged children. Those who participated in Follow Through as model-sponsors believe that change cannot come from within the local-schools on any major scale and that incentives for change more powerful than those presently available must be provided. The paradox is apparent - the literature on change recognizes that change must be desired by those within the system, but experience reveals that those within the system who want to change require more support and knowledge than can possibly come from within. An immediate solution to this paradox is not apparent. Model-sponsorship is only one possible avenue. (p. 74)

The model sponsors stress the need to understand complex relationships within a community as well as between a community and an educational program. Thus the sponsors have recommended that 1) "Information on the status of a school system prior to the initiation of an intervention is needed," and 2) "The data demonstrate that communities differ radically. More information is needed on how to identify and index these differences to determine how they affect the implementation of a model." We also want to note the sponsors have recommended that 3) "State education agencies should be involved in educational changes in the schools at a more meaningful level than that mandated by current Follow Through regulations," and 4) "Federal

government decision making concerning large-scale programs like Follow Through must become more timely and better coordinated.

We agree that with the suggestion that more attention to the particular conditions within a community is likely to result in a richer description of an educational program and thus lead to a more useful kind of knowledge. But we sense that there is more to the sponsors' position as stated. There seems to be the implication that the variety occurring within communities can be characterized by an underlying model, rather than accumulated experience. Assuming that educational settings conform to this model in a systematic fashion, the problem is to establish enough of the theory to improve educational practice. Moreover, we hear a tacit approval of large-scale educational evaluations with the corresponding "assistance" of federal and state agencies. "Stronger incentives" to change from the outside can also become a means of unwanted intervention, or worse, imposition.

2. Experimental Design & Statistics.

Our low estimation of the importance of quantitative experimental methods in the evaluation of Follow Through is not a reflection of a more general attitude of suspicion about the validity of these methods. With the exception of the concern lavished on statistical inference* by USOE/SRI/ABT evaluators, the methods of design and analysis used in the past were generally adequate to the purpose; the purpose is now outdated and inappropriate.

---

*Inferential statistical concerns have been over-emphasized in past FT evaluations. Alpha levels make sense either when based on explicit probabilistic sampling (as in well-done surveys) or when based on randomization in experiments (thus providing a permutation interpretation of alpha levels). Without either (the situation in Follow Through evaluation), statistical inference means little; it merely gives a false sense of confidence in the findings and draws attention from the more complex questions of generalization that statistics will not solve.

There is nothing inherently inadequate in quasi-experimental methodology; indeed, when handled well it is an impressive collection of tools (Cook and Campbell, 1979). Much criticism has been leveled against the analysis of covariance (ANCOVA), and yet many statistical studies of bias have shown covariance adjustments to be among the best (Rubin, 1973; Cochran and Rubin 1973). Different methods of correcting for fallible covariates provide plausible bounds for a true estimate of treatment effects. Nevertheless, we will rehash briefly the issue of fallible covariates; we will consider the case of a single covariate (with multiple covariates, a single "best" linear combination can be formed).

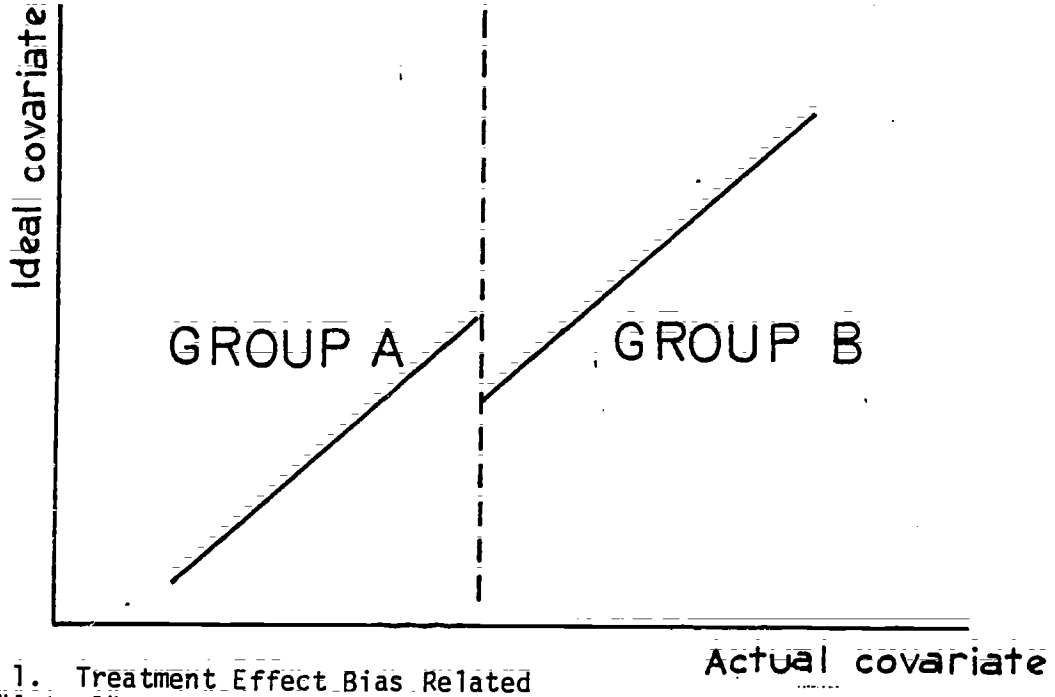

Figure 1. Treatment Effect Bias Related to "Actual" and "Ideal" Covariates

As illustrated in Figure 1, bias can be characterized by a discontinuity in a linear function relating the ideal to the actual covariate. "Reliability" corrections may increase the slope of the within-groups regressions and thus reduce the discontinuity. An underestimate of reliability, though, increases the slopes too much resulting in an overcorrection. One is left with the

problem of determining the appropriate type of reliability. Internal consistency, stability or generalizability coefficients do not insure the relevance (Cronbach, et al. 1977) of the covariate. One may do better to think of the covariate as a linear model of the rules or process by which individuals were assigned to groups (e.g., Follow Through and non-Follow Through). The more accurately the ideal assignment rule is approximated, the more accurate the covariance adjustment (Cronbach, et al., 1977). Discriminant analysis can be used to assess assignment rules in terms of errors of misclassification. When this analysis was applied to the data from the Abt Follow Through evaluation, it was found that pupils could be assigned to their appropriate Follow Through or non-Follow Through (control) group with an error rate of only 20% (using the observed covariates in the discriminant function to make the classification) (Camilli, 1980). Considering the guidelines for selecting Follow Through pupils, this error rate is suprisingly low. Thus, there is evidence that covariance adjustments of the kind applied in past Follow Through evaluations may be adequate. But for a decent quantitative experimental evaluation to be possible, adequate measures of program success are needed. Here, past efforts broke down. First some technical problems, then the tough problems.

In the Abt Follow Through evaluation the logistics of mass testing overwhelmed the effort. The actual testing was not standardized across sites and models; moreover, testing itself was valued differently in different places and by different persons. The result was some interesting features in the data that heretofore have not been sufficiently noticed nor commented upon. For example, for non-Follow Through pupils, the percentage

of the pupils scoring below chance on some Metropolitan Achievement Test (MAT) subtests varied from 5% to 35% across the models. Also, in Follow Through groups, relatively large proportions of children scored zero on multiple-choice subtests of 35 items, four options per item. Even more interesting, perhaps, is the relation of percentage of non-Follow Through model gains. In Figure 2, although only 14 models are plotted, a strong positive relationship is observed for MAT reading. Furthermore, when percent of Follow Through pupils below chance is partialed out, the strength of the relationship does not decrease (Camilli, 1980). Thus, the invalidity of the testing procedure is evident in the test scores themselves.

In going so far to comment on these technical issues, we risk creating the impression that we believe that the key questions about Follow Through evaluation are technical or amenable to technical solutions. We do not. The important problems with Follow Through evaluation are not technical. They will not be solved with Rasch models or factor analysis or "principal-axis adjustments" (which Wisler, Burns and Iwamoto, 1978, believe would be a useful addition to future Follow Through evaluations). The problems will not be solved by eliminating test scores below chance levels. Testing itself is not valued equally in different Follow Through models. Some see it as an obnoxious intrusion; others drill pupils for weeks on item-forms. Test results must not be the sole or even primary indicators of success. To invest large amounts of money in attempts to synthesize test results for "education managers" is indefensible. Mass testing is a growing federal tendency. It exists primarily as an attempt to simplify

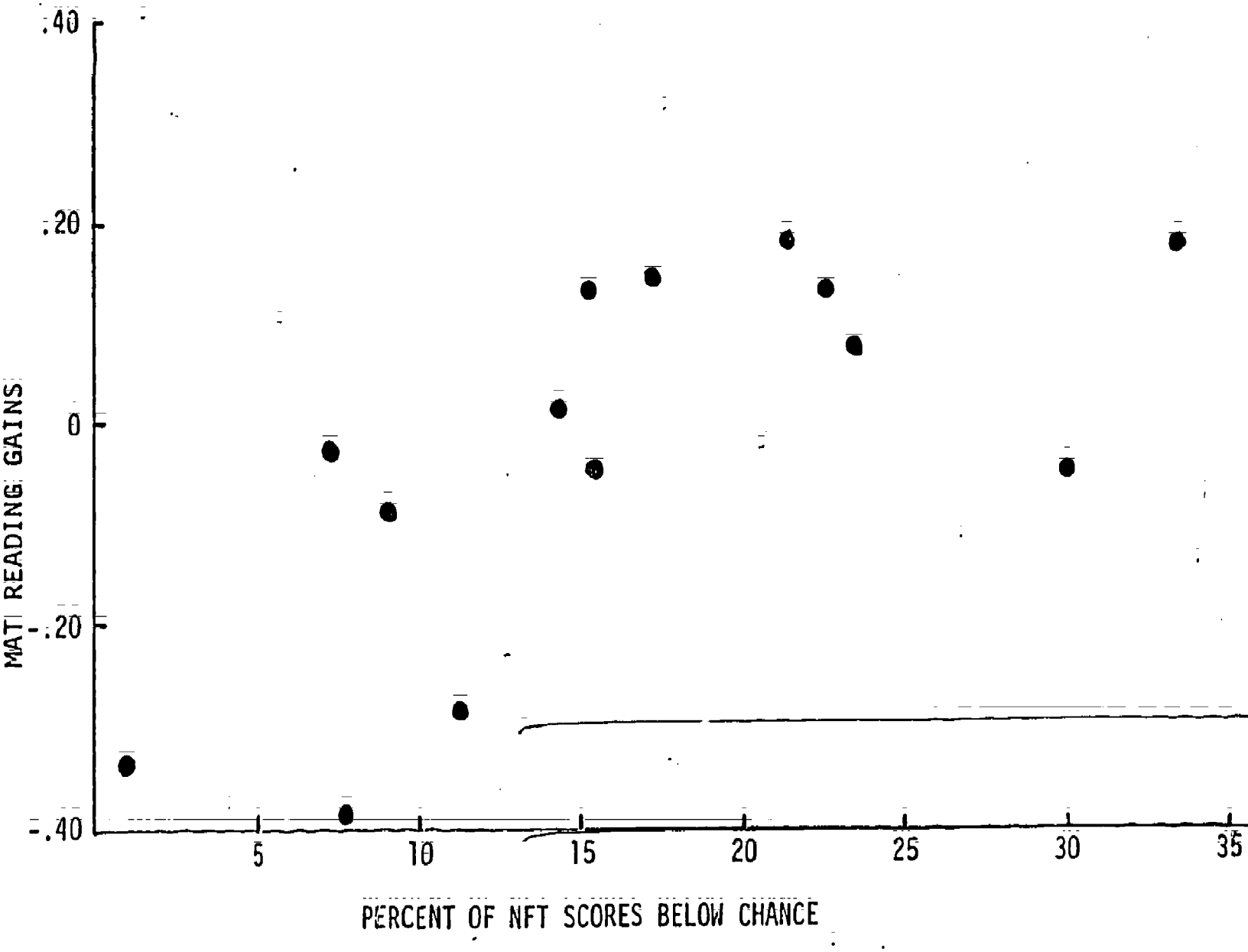

Figure 2. Plot of treatment effects against percentage of NFT reading scores below chance.

intrinsically complex issues.

The art of teaching must not be subordinated to the technology of mass testing. These tests homogenize educational goals and aspirations -- an ironic fate for a program founded on the slogan "planned variation."

3. Discovering New Knowledge About Follow-Through.

"New" Knowledge. There is much rhetoric in NIE prospectuses for Follow Through evaluation about "new models" of early compensatory education. We see nothing on the horizon that would justify this optimistic expectation. We suspect that the rhetoric derives either from vain hopes or the need of NIE to justify its 20% of Follow Through funds in terms of its mission, viz., research, not service.

To anyone but a gullible professional educator caught up in enthusiasm for one gimmic: or another, the idea that there are even 22 "models" of early compensatory education is ludicrous. Indeed, these 22 exhaust (with much redundancy) the pedagogic imagination: Pestalozzi via Montessori; Freud via Bank Street; Watson via Skinner; not to mention Jesus Christ via Thomas Aquinas (since Follow Through is a public project). Does anyone really believe that there is something all that remarkable hiding in the bushes that will be discovered by federal evaluation programs? NIE planning documents refer to "media," and "home learning," and the new problems attendant on the rise of "single-parent families." This all sounds quite up-to-date; but none of it rings true. We do not doubt in the least that a dozen educators can be found who can generate excitement about the prospects of some psycho-linguistic trick or the promise of "cognitive psychology." But we doubt seriously that an educator can be found who is capable of improving on the variegated state of the art

already represented by the two dozen extant Follow Through models. We don't need more Follow Through models; we need more thinking and less measuring on the models that already exist. A brilliant young psychoanalyst and Marxist in Berlin in the 1920's, Siegfried Bernfeld, commented as follows on the historical fads that swept the educational world of early twentieth century Europe:

"All of the educational means thought to be appropriate for changing the child's naive and intuitive personality into some higher form are suspiciously simple and trite. Jointly and individually, they are not new; thus is their banality revealed. Nor is it likely possible to devise a genuinely new method of education. Certainly the great pedagogues have not succeeded. It matters little whether today they recommend the power of love or strict discipline, whether they recommend teaching by word or by example or by the rod, whether they demand the teacher's active involvement or his patient attendance, whether they insist on the rechanneled acting out of the child's impulses or their repression. Ever since there were parents and teachers, the ancient gamut from a mere stern glance to prison punishment have all been tried. Children came to be the multitude by motley combinations of these methods, and more multitudes were raised by this multitude; there can be no combinations that have not already been tried -- and the result is the mankind of today, of any day...the banal and commonplace methods possess none of the power to transform and perfect, which the great pedagogues ascribe to them. There is no magic, neither in the teacher's gentle rebuke nor in the salutary thrashing." (Bernfeld, 1928, p. 38; translation by the first author.)

Knowledge and Human Interests. Many planners of evaluations are curiously inconsistent in the way they treat potential audiences for their findings and the way they treat themselfes. Those who argue vigorously for the superiority of quantitative experimental evaluation (like that which characterized Title I and Follow Through evaluation in USOE) fail to realize that their own favorable evaluation of this method is not itself based on evidence from quantitative experimental evaluations. They cannot point to experiments that show quantitative experimental evaluations to be superior to other kinds (e.g., to ethnographic/case-study evaluations, evaluations by personal experience, evaluations of other than cost-benefit issues); rather they judge the superiority of their chosen methods on the basis of personal experience, loosely analogous experience, historical analyses and other non-quantitative, non-experimental ways of knowing and evaluating. It would be inappropriate to pursue an investigation of personal motives too far, but it seems safe to say that some people's preference for quantitative experimental evaluation is little more than an expression of personal taste; they fear the future, they fear ambiguity, and they like a neat desk, tidy plans and simple answers. And yet, when it comes to the question of other persons' right to decide how to teach children, these same evaluators deny others the right to decide on grounds epistemologically equivalent to those on which the evaluators choose their evaluation methods. If it were not that so many people were intimidated by the evaluators' methods, their arbitrary authority would more quickly be seen as illegitimate.

Consider the following situation. You have before you two pieces

of information about a Follow Through model: a) an experimental study showing that this particular model produces average achievement .10 standard deviations higher than conventional teaching; b) a complete descriptive study of the model's goals, aspirations and procedures, the feelings and reactions of teachers and parents who have worked with the model for five years, a critique of the model by educational philosophers and methodologists who have seen the model work, and other pertinent observations of a resident observer of the model which we cannot list since they cannot be anticipated in detail. Which of the two reports will you accord greater weight? Before answering, bear this in mind. The Church of Scientology and the Maharishi International University (Fairfield, Iowa, in case you think we made it up) can cite experimental studies validating their approach that show effect sizes bigger than the .10 effects typical of the impact of the Follow Through models on achievement measures (Ferguson, 1980). Yet, are you in personal danger of enrolling soon in either Scientology or MIU? We doubt it. "Well," you say, "the experimental effects of Transcendental Meditation are only a small part of what is entailed in a decision to matriculate at MIU." "Indeed, even as many people feel that Metropolitan Achievement Test scores are only a small part of judging Follow Through models," we respond. When we confront a significant choice (career, marriage, family, political party, friendships and the like), we worry about what will be expected of us as persons and how that accords with our feelings about our integrity, our contribution to our loved ones and friends, our happiness and our moral obligations. We worry little about .10 effects on measurable variables.

The audience for Follow Through evaluation is those professional educators in schools who worry about teaching young poor children.

They are not interested in evaluators' experiments nor education managers edits. They decide how to teach on the basis of enormously complex and partly private attractions and antipathies. And, dear reader, if you feel inclined to scorn their unscientific and irrational minds, reflect again on the fact that you do not want your child to enroll at Maharishi International University or that the Scarsdale diet appeals to you far more than hypnosis as the way to lose 20 pounds because the idea of hypnosis strikes you as weird and unsettling.

The audience for Follow Through evaluation wants to know much more than experiments and measurements can tell. They want to know what is expected of teachers who use this Follow Through model. Is it consistent with their view of themselves as professionals, as saviors of poor children, as "instructional managers"? Does this model treat pupils as though they were robots, or delicate flowers, or children of God? If they, the teachers, adopt the prescribed role will they grow to be like Jean Piaget, or Maria Montessori, or Anna Freud, or Siggy Engelmann? Are teachers treated as intelligent human beings or merely as means toward technically prescribed ends and instruments of someone else's will? What do people really want to know about Direct Instruction, to pick an example? They don't care whether DI can coach pupils to spell more words correctly than can Bank Street. They want to know if there is any substance to the rumors that DI is psychological torture for the children who go through it or if DI teachers grow to feel demeaned and superfluous. One does not answer these questions adequately by asking Becker and Engleman nor by administering the Metropolitan Achievement Test. Likewise,

people want to know what kind of personalities their children would be exposed to if they were enrolled in a Bank Street program, with its mildly unsettling armona of Freudianism. Such are peoples' concerns. They can not be discounted or ignored on grounds of pedagogic efficiency, cost-effectiveness, democratic decision-making nor the rational conduct of public affairs, each of which is a value honored by those who presume to evaluate Follow Through.

Worthwhile Knowledge About Education. In complex evolutionary systems like education, it is generally more important to evaluate an image of the future than to evaluate current accomplishments (Boulding, 1978). The key perception of value may be nearer to the recognition of potential than the confirmation of current productivity. Educational technologists are fond of pointing out correctly that the steam engine lsot its first race with a horse. How in 1970, should one have judged the value of stock in Post Slide-rule Company vs. Texas Instrument? It is the dependence of valuations of educational enterprises on their images of the future and the low predictability of these futures that make educational evaluation such a risky business (Glass, 1979).

HOW SHOULD NIE EVALUATE FOLLOW THROUGH?

House (1980) criticized the model of evaluation that grew up during the 1970's in USOE and now threatens to infect NIE's efforts to evaluate Follow Through:

> "Federal evaluation policy has been based on the systems analysis approach. Its major audiences are managers and economists. It assumes consensus on goals, on known cause and effect, and on a few quantified outcome variables. Its methodology includes planned variation experimentation and cost-benefit analysis. Its end is efficiency. It asks the question, What are the most efficient programs?

"As Articulated by major proponents like Rivlin and Evans, it assumes there is a direct parallel between the production of social services and manufacturing. The same analysis techniques will apply. The only true knowledge is a production function specifying stable relationships between input and output. The only way to such knowledge is through experimental methods and statistical techniques. It is possible to agree on a few output measures. The issue is efficient allocation of resources.

"The key decisions will be made at higher government levels, and tough management can do the job. The ultimate justification is utilitarian -- to maximize satisfaction in society. To maximize, one must know which programs are most efficient. This can be done only by comparing alternatives, for which one must have a common measure of output. This is a job for experts.

"There are places where this approach can be applied successfully. But the United States as a whole is not one of them. The approach can be successfully applied where there really are only a few goals and outcome measures. This is likely to happen where the audience for the evaluation is very narrowly defined and agrees on a few criteria of comparison. It also helps if the criteria can be represented by a reasonably valid quantitative indicator." (House, 1980, p. 222)

The USOE evaluation of Follow Through took ten years and cost $20 million; it was not worth the money. And those who were primarily responsible for its form (wherever they are today) remain doggedly unrepentant.

"The identification of successful sites, combined with the often weak or variable model effects, suggests that local conditions, such as children whose needs match especially well what the model can provide, local variations of the model, or especially skilled teachers, were more apt to determine success than the models used. We do not think this means that work on educational models like those implemented in Follow Through should be abandoned. A few models had results consistent enough to warrant continuing development and testing of these and other models. It is possible that more models might have shown positive results if they had been more precisely specified at the outset and more faithfully and uniformly implemented in the school setting.

"For this and other reasons, we think that Follow Through has not been a fair test of whether or not we can learn from a large-scale educational experiment. Launched hastily because of an

> unexpected turn in congressional appropriations, the Follow Through experiment never really righted itself. Nonetheless, because of the accountability movement in education, the potential for running sound experiments may be even better today than it was in 1968." (Wisler, Burns & Iwamoto, 1978, p. 180)

> "... we disagree that there should be less government control of evaluations generally. The early difficulties with the Follow Through evaluation can be traced to lack of sound and strong direction from OE, not to interference from the government. The evaluation was salvaged only after Garry McDaniels and others assumed control of it in 1971. Our experience with other federal education evaluations suggests that weak direction from OE is a sure guarantee of a useless evaluation. We also disagree with the conclusion that the type of evaluation used for Follow Through is no longer needed." (Wisler, Burns & Iwamoto, 1978, p. 179)

The course on which NIE evaluation of Follow Through is set (or will soon be set) threatens to honor, unwittingly perhaps, the values of "science" as they are viewed by logical positivists (particularly behavioral psychogists, who almost alone among observers are pleased with the results of past attempts to evaluate Follow Through). These values are described by their friends as "objectivity through operationism" and by their enemies as "Fliegenbeinenzahlen" (literally "counting flies legs" or figuratively, as trivializing overquantification). In a democracy, there is a great range of values that must be honored by those who presume to evaluate in the public interest; and those values go far beyond what is now measureable I am reffering to such things as dignity, respect and love. And the thought that these are merely multivariate outcome variables that will yield their secrets to the scientific coaxing of factor analysis is a thought hopelessly held prisoner by the shackles of logical positivism.

This is a disheartening future for Follow Through evaluation if it is indeed the future that NIE is in danger of bringing about. But it is precisely what is to be expected because NIE is going back to the same experts who gave USOE the same old advice about measurement, design and analysis (only now the advice is proped up with false hope and excuses for past

failures).

The convergence of past Follow Through evaluations on the common, easily measurable outcomes is having the unwholesome effect of homogenizing the evolution of programs. In education as elsewhere, the old adage holds true: enemies grow to resemble each other. Where organizations and people fight in a zero-sum battle for the same resources, in time they grow increasingly alike. Thus the damage wrought by evaluation on a few criteria that are currently prepared for mass testing is doubly serious. It is not only unfair to contemporary efforts whose benefits are poorly understood, but it warps the evolution of efforts that might otherwise have made unanticipated accomplishments.

NIE is too dangerously close to believing the history that USOE writes about its evaluation experiences. Already the NIE plans for Follow Through evaluation smack of the USOE model. Quoting from the October 1, 1980, "Plans for Follow Through Research and Development":

> "As one cohort of approaches is fully tested, it will be phased out of funding, results will be disseminated, and another cohort of approaches will be phased in. Through this strategy, it is planned to continually infuse the Follow Through Program with new research-based knowledge to improve its effectiveness." (p. 7)
>
> "... NIE will test a small number of approaches to school improvement in the management and implementation area and document their effectiveness with sufficient detail so that the results are replicable for widespread dissemination in Follow Through and elsewhere." (p. 12)

The conception of evaluation that seems to underlie the NIE planning document does not accord with the reality of how schools change or how educators create and grow. And worse yet, this reality is increasingly difficult to discern because the federal government is changing schools to accord with its own image of how knowledge should be produced,

disseminated and used. By controlling Dissemination Panels and the "validating" of programs and money to induce schools to join a system of knowledge production and use, the federal government risks changing schooling into the image of its own conceptualization and risks the loss of value in a broader and truer sense. Planning and control tend to create self-confirming futures and destroy alternative futures; alternatives (variations) are essential to growth and change.

How should NIE evaluate Follow-Through? Like it has never before been evaluated.

1. NIE should dispense with the fiction that the purpose of Follow Through evaluation is to validate and invalidate models. Indeed, it should admit that the continued existence of approaches to teaching poor children does not depend on government-sponsored field experiments.
2. NIE should disabuse itself of the myth that "new models" are likely to be "discovered" by any methods, particularly by the methods of quantitative experimental design.
3. Instead of imitating past efforts, the NIE should conduct evaluation that emphasizes description (principally qualitative) for informed choice. Models should be described in terms that people consider personally significant when they choose a particular profession for themselves or a school for their children. (By contrast, the language of current NIE planning documents is technocratic, behavioristic and anti-democratic.) An ethnographic or case-study approach

to evaluation should be adopted in place of a quantitative, experimental field trial. What one needs to know about Follow Through models is not more statistics (these exist in abundance) but rather

a) coherent, detailed portrayals of life in school for pupils, teachers and parents as it is colored and shaped by allegiance to a particular Follow Through model,

b) such portrayals having been written by disinterested, expert ethnographers with at least two years on-site for data collection and,

c) such portrayals being focused on a broad range of concerns including the model's philosophy, its history (since its future must be projected), techniques, financial and psychic-costs, side-effects and after-effects, the roles it requires people to play, its potential for a favorable evolution, and the like.

SUMMARY

Past evaluations of Follow Through were quantitative and experimental. They created [redacted] dissent and changed few minds. "Models" of compensatory education are minor influences in pupils' development. More important in children's growth are their native endowment, their health, how their parents and siblings treat them, and other influences not controlled by schools.

The deficiencies of quantitative, experimental evaluation are thorough and irreparable. The problem lies less with experimental designs for assessing causal impact than with the impossibility of translating complex, subtle and vague notions of child development and education into tests for mass administration.

There are probably at most a half-dozen genuinely and importantly different approaches to teaching children and these are already well-represented in existing Follow-Through models.

The audience for Follow-Through evaluations is an audience of teachers. This audience does not need the statistical findings of experiments when deciding how best to educate children. They decide such matters on the basis of complicated public and private understandings, beliefs, motives and wishes. They have the right and good reasons so to decide.

The course on which NIE evaluation of Follow Through is set threatens to honor, unwittingly perhaps, the values of "science" as they are viewed by logical positivists, mainly behavioral psychologists. There is a greater range of values that in a democracy must be honored by those who presume to evaluate in the public interest; and those values go far beyond what is now measurable.

NIE should dispense with the fiction that the purpose of Follow-Through evaluation is to validate and invalidate models. Indeed, it should admit that the continued existence of approaches to teaching poor children does not depend on government-sponsored field experiments.

NIE should disabuse itself of the myth that "new models" are likely to be "discovered" by any methods, particularly by the methods of quantitative experimental design.

In place of past efforts, the NIE should conduct evaluation that emphasizes description (principally qualitative) for informed choice. Models should be described in terms that people consider personally significant when they choose a particular profession for themselves or a school for their children. (By contrast, the language of current NIE planning documents is technocratic and anti-democratic.) An ethnographic or case-study approach to evaluation should be adopted in place of a quantitative, experimental field trial. What one needs now is not more statistics but rather

a) coherent, detailed portrayals of life in school for pupils, teachers and parents as it is colored and shaped by allegiance to a particular Follow Through model,

b) written by disinterested, expert enthnographers with at least two years on-site for data collection and,

c) focused on a broad range of concerns including the model's philosophy, its history (since its future must be projected), techniques, financial and psychic-costs, side-effects and after-effects, the roles it requires people to play, its potential for a favorable evolution, and the like.

REFERENCES

Bereiter, C. and Kurland, M. Where some Follow Through models more effective than others? Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 30 March 1978.

Bernfeld, S. Sisyphos, Oder die Grenzen der Erziehung. Leipzig: Internationaler Psychoanalytischer Verlag, 1928.

Boulding, K. E. Ecodynamics: A New Theory of Societal Evolution. Beverly Hills, Calif.: SAGE Publications, 1978.

Camilli, G. A. A reanalysis of the effect of Follow Through on cognitive and affective development. Unpublished doctoral dissertation, University of Colorado, 1980.

Cochran, W. G. and Rubin, D. B. Controlling bias in observational studies: a review. Sankhyā, 1973, 35, 417-446.

Cook, T. D. and Campbell, D. T. Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand-McNally, 1979.

Cronbach, L. J. et al. Analysis of covariance in nonrandomized experiments: parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium, Stanford University, 1977.

Ferguson, P. C. An integrative meta-analysis of psychological studies investigating the treatment outcomes of meditation techniques. Doctoral dissertation, University of Colorado, 1980.

Glass, G. V Policy for the unpredictable: uncertainty research and policy. Educational Researcher, October, 1979, 8, 12-14.

Hodges, W. et al. Forces for Changes in the Primary Schools. Ypsilanti, Mich.: Hi/Scope Press, 1980.

House, E. R. Evaluating with Validity. Beverly Hills, Calif.: SAGE Publications, 1980.

House, E. R.; Glass, G. V; McLean, L. D.; and Walker, D. No simple answer: critique of the Follow Through evaluation. Harvard Educational Review, 1978, 48, 128-160.

Rubin, D. B. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics, 1973, 29, 184-203.

Stebbins, L. B. et al. Education as Experimentation: A Planned Variation Model. Vol. 4-A: An Evaluation of Follow Through. Cambridge, Mass.: Abt Associates, 1977.

Wissler, C. E.: Burns, G. P.; and Iwamoto, D. Follow Through redux: a response to the critique by House, Glass, McLean, and Walker. Harvard Educational Review, 1978, 48, 171-185.

APPENDIX

In their new book Toward Reform of Program Evaluation, Cronbach and his associates (1980) listed 95 theses about the proper roles, methods and uses of evaluation. Although they invited readers to discuss the theses and sharpen their thinking on them, there is no mistaking the fact that these declarative statements represent the results of the group's best thinking about evaluation. And that thinking is remarkably broad and perspicacious. Moreover, the theses provide an excellent background against which to critique the thinking on evaluation that characterized USOE efforts in Title I and Follow Through evaluation in recent years.

In the list that follows, we have marked with an asterisk each assertion that clearly runs counter to the federal model of program evaluation that came to characterize USOE and threatens to influence NIE.

Ninety-Five Theses

1. Program evaluation is a process by which society learns about itself.
X 2. Program evaluations should contribute to enlightened discussion of alternative plans for social action.
3. Evaluation is a handmaiden to gradualism; it is both conservative and committed to change.
X 4. An evaluation of a particular program is only an episode in the continuing evolution of thought about a problem area.
5. The better and the more widely the workings of social programs are understood, the more rapidly policy will evolve

and the more the programs will contribute to a better quality of life.

6. Observations of social programs require a closer analysis than a lay interpreter can make, for unassisted judgment leads all too easily to false interpretations.
7. In debates over controversial programs, liars figure and figures often lie; the evaluator has a responsibility to protect his clients from both types of deception.

* * *

8. Ideally, every evaluation will inform the social system and improve its operations, but everyone agrees that evaluation is not rendering the service it should.
9. Commissioners of evaluations complain that the messages from evaluations are not useful, while evaluators complain that the messages are not used.

* * *

10. The evaluator has political influence even when he does not aspire to it.

X 11. A theory of evaluation must be as much a theory of political interaction as it is a theory of how to determine facts.

X 12. The hope that an evaluation will provide unequivocal answers, convincing enough to extinguish controversy about the merits of a social program, is certain to be disappointed.

X 13. The evaluators' professional conclusions cannot substitute for the political process.

14. The distinction between evaluation and policy research is disappearing.

* * *

X 15. Accountability emphasizes looking back in order to assign praise or blame; evaluation is better used to understand events and processes for the sake of guiding future activities.

16. Social renovations disappoint even their architects.
17. Time and again, political passion has been a driving spirit behind a call for rational analysis.

X 18. A demand for accountability is a sign of pathology in the political system.

* * *

X 19. An open society becomes a closed society when only the officials know what is going on. Insofar as information is a source of power, evaluations carried out to inform a policy maker have a disenfranchising effect.

X 20. The ideal of efficiency in government is in tension with the ideal of democratic participation; rationalism is dangerously close to totalitarianism.

* * *

X 21. The notion of the evaluator as a superman who will make all social choices easy and all programs efficient, turning public management into a technology, is a pipe dream.

X 22. A context of command, with a manager in firm control, has been assumed in nearly all previous theories of evaluation.

23. An image of pluralistic accommodation more truly represents how policy and programs are shaped than does the Platonic image of concentrated power and responsibility.

24. The evaluator must learn to serve in contexts of accomodation and not dream idly of serving a Platonic guardian.

X 25. In a context of accommodation, the evaluator cannot expect a "go/no-go" decision to turn on his assessment of outcomes.

* * *

X 26. What is needed is information that supports negotiation rather than information calculated to point out the "correct" decision.

27. Events move forward by piecemeal adaptations.

X 28. It can scarcely be said that decisions about typical programs are "made"; rather, they emerge.

X 29. The policy-shaping community does not wait for a sure winner; it must act in the face of uncertainty, settling on plausible actions that are politically acceptable.

* * *

30. It is unwise for evaluation to focus on whether a project has "attained its goals."

X 31. Goals are a necessary part of political rhetoric, but all social programs, even supposedly targeted ones, have broad aims.

32. Legislators who have sophisticated reasons for keeping goal statements lofty and nebulous unblushingly ask program administrators to state explicit goals.

X 33. Unfortunately, whatever the evaluator decides to measure tends to become a primary goal of program operators.

* * *

34. Evaluators are not encouraged to ask the most trenchant questions about entrenched programs.
35. "Evaluate this program" is often a vague charge because a program or a system frequently has no clear boundaries.
X 36. Before the evaluator can plan data collection, he must find out a great deal about the project as it exists and as it is conceived.
X 37. A good evaluative question invites a differentiated answer instead of leaving the program plan, the delivery of the program, and the response of clients as unexamined elements within a closed black box.
X 38. Strictly honest data collection can generate a misleading picture unless questions are framed to expose both the facts useful to partisans of the program and the facts useful to its critics.

* * *

X 39. Before laying out a design, the evaluator should do considerable homework. Pertinent questions should be identified by examining the history of similar programs, the related social theory, and the expectations of program advocates, critics, and prospective clients.
40. Precise assessment of outcomes is sensible only after thorough pilot work has pinned down a highly appropriate form for an innovation under test.
X 41. When a prototype program is evaluated, the full range of realizations likely to occur in practice should be observed.
42. Flexibility and diversity are preferable to the rigidity written into many evaluation contracts.

* * *

43. The evaluator who does not press for productive assignments and the freedom to carry them out takes the King's shilling for selfish reasons.
44. The evaluator's aspiration to benefit the larger community has to be reconciled—sometimes painfully—with commitments to a sponsor and to informants, with the evaluator's political convictions, and with his desire to stay in business.
45. Managers have many reasons for wishing to maintain control over evaluative information; the evaluator can respect all such reasons that fall within the sphere of management.
46. The crucial ethical problem appears to be freedom to communicate during and after the study, subject to legitimate concerns for privacy, national security, and faithfulness to contractual commitments.
47. With some hesitation, we advise the evaluator to release findings piecemeal and informally to the audiences that need them. The impotence that comes with delay may be a greater risk than the possibility that early returns will be misread.

* * *

48. Nothing makes a larger difference in the use of evaluations than the personal factor—the interest of officials in learning from the evaluation and the desire of the evaluator to get attention for what he knows.
49. Communication overload is a common fault; many an evaluation is reported with self-defeating thoroughness.

X 50. Much of the most significant communication of findings is informal, and not all of it is deliberate; some of the most significant effects are indirect, affecting audiences far removed from the program under investigation.
51. An evaluation of a particular project has its greatest implications for projects that will be put in place in the future.
52. A program evaluation that gets attention is likely to affect the prevailing view of social purposes, whether or not it immediately affects the fate of the program studied.
X 53. Advice on evaluation typically speaks of an investigation as a stand-alone study that will draw its conclusions about a program in complete isolation from other sources of information.
X 54. It is better for an evaluative inquiry to launch a small fleet of studies than to put all its resources into a single approach.
X 55. Much that is written on evaluation recommends some one "scientifically rigorous" plan. Evaluations should, however, take many forms, and less rigorous approaches have value in many circumstances.
X 56. Results of a program evaluation are so dependent on the setting that replication is only a figure of speech; the evaluator is essentially an historian.

* * *

X 57. An elegant study provides dangerously convincing evidence when it seems to answer a question that it did not in fact squarely address.
X 58. Merit lies not in form of inquiry but in relevance of information. The context of command or accommodation, the stage of program maturity, and the closeness of the evaluator to the probable users should all affect the style of an evaluation.
X 59. The evaluator will be wise not to declare allegiance to either a quantitative-scientific-summative methodology or a qualitative-naturalistic-descriptive methodology.
60. External validity—that is, the validity of inferences that go beyond the data—is the crux; increasing internal validity by elegant design often reduces relevance.

---

* * *

X 61. Adding a control costs something in dollars, in attention, and perhaps in quality of data; a control that fortifies the study in one respect is likely to weaken it in another.

62. A strictly representative sample may provide less information than a sample that overrepresents exceptional cases and deliberately varies realizations.

X 63. The symmetric, nonsequential designs familiar from laboratory research and survey research are rarely appropriate for evaluations.

64. Multiple indicators of outcomes reinforce one another logically as well as statistically. This is true for measures of adequacy of program implementation as well as for measures of changes in client behavior.

* * *

65. In project-by-project evaluation, each study analyzes a spoonful dipped from a sea of uncertainties.

X 66. In any primary statistical investigation, analyses by independent teams should be made before the report is distributed.

X 67. Evaluations of a program conducted in parallel by different teams can capitalize on disparate perspectives and technical skills.

X 68. The evaluator should allocate investigative resources by considering four criteria simultaneously: prior uncertainty about a question, costs of information, anticipated information yield, and leverage of the information on subsequent thinking and action.

69. A particular control is warranted if it can be installed at reasonable costs and if, in the absence of that control, a positive effect could be persuasively explained away.

70. The importance of comparative data depends on the nature of the comparison proposed and on the stage of program maturity.

X 71. When programs have multiple and perhaps dissimilar outcomes, comparison is invariably judgmental. No technology for comparing benefits will silence partisan discord.

* * *

72. Present institutional arrangements for evaluation make it difficult or impossible to carry on the most useful kinds of evaluation.

73. In typical federal contracting, many basic research decisions are made without consulting the evaluators who will do the work.

X 74. The personal scientific responsibility found in ordinary research grants is lacking in contract evaluation; the "principal investigator" is a firm with interchangeable personnel.

75. Though the information from an evaluation is typically not used at a foreseeable moment to make a foreseen choice, in many evaluations a deadline set at the start of the study dominates the effort.

76. Evaluation contracts are increasing in size, but tying many strands into a single knot is rarely the best way to get useful information.

X 77. Large-scale evaluations are not necessarily better than smaller ones.

X 78. Major evaluations should have multiple sponsorship by agencies with different perspectives.

X 79. Decentralizing much evaluation to the state level would be a healthy development.

* * *

80. Society will obtain the assistance that evaluations can give only when there is a strong evaluation profession, clear about its social role and the nature of its work.

81. There is a boom town excitement in the evaluation community, but in constant dollars federal funding for evaluation research has regressed in the last few years.

82. It is inconceivable that evaluators will win their battle for appropriate responsibilities if they remain unacquainted with one another, insensitive to their common interests, and fractionated intellectually.

* * *

83. For any suitably broad social problem, a "social problem study group" should be set up. It would be charged to inform itself by weighing, digesting, and interpreting what is known. It would foster needed investigations and make the policy-shaping community aware of what is and is not known.

* * *

84. Honesty and balance in program evaluation will be increased by critical review of the performance of evaluators and sponsors.
85. Oversight by peers is the most promising means of upholding professional standards and of precipitating debate about strategic and tactical issues.
86. The best safeguard against prematurely frozen standards for evaluative practice is multiple, independent sources of criticism.
87. There is need for exchanges more energetic than the typical academic discussion and more responsible than debate among partisans.
88. Reviews of evaluation should be far more frequent than at present, and reviews from diverse perspectives should appear together.

* * *

89. For the prospective evaluator, basic training at the doctoral level in a specific social science is preferable to training restricted to evaluation methods.
90. Training in evaluation is too often the stepchild of a department chiefly engaged in training academicians or providers of service.
91. Case-study seminars scrutinizing diverse evaluative studies provide a needed interdisciplinary perspective.

92. Internships with policy agencies that use evaluation sensitize future evaluators to the realities of evaluation use and nonuse. These realities are hard to convey in a classroom.

* * *

93. The evaluator is an educator; his success is to be judged by what others learn.
X 94. Those who shape policy should reach decisions with their eyes open; it is the evaluator's task to illuminate the situation, not to dictate the decision.
X 95. Scientific quality is not the principal standard; an evaluation should aim to be comprehensible, correct and complete, and credible to partisans on all sides.